Ordering Slots of Semantically Related Schemata of Polish Verbs

Elżbieta Hajnicz

Institute of Computer Science, Polish Academy of Sciences
Elzbieta.Hajnicz@ipipan.waw.pl

Abstract

In this paper a method of ordering slots of verb schemata related by diathesis alternations is presented. Such schemata are grouped together and represent a single meaning of a verb. The schema dominating in a group is found and its slots are numbered w.r.t. their obliqueness hierarchy. These numbers are propagated to other schemata accordingly to alternations linking slots of schemata related by them.

1. Introduction

The primary task of our research is to create a semantic valence dictionary in an automatic way. To accomplish this goal, the valence dictionary of Polish verbs is supplemented with semantic information, provided by wordnet's semantic categories (Hajnicz, 2009c) or synsets (Hajnicz, 2009a) of nouns. In our present work we focus on slots being nominal phrases and prepositional-nominal phrases, whose semantic heads are nouns. We discuss the case of 25 predefined semantic categories of nouns, which is simpler than the case of actual wordnet synsets.

In our previous works (Hajnicz, 2009b, 2010) we focused on the preparation of a syntactic-semantic valence dictionary, in which each slot of a syntactic schema[1] is supplied with a list of semantic categories, forming a semantic frame. However, a genuine semantic dictionary is composed of semantic frames represented as a predicate-argument structure, in spite of their syntactic realisations. Each semantic argument is connected with its semantic role. On the other hand, syntactic slots of schemata are provided with information concerning which semantic role they realise.

A method of grouping schemata that participate in diathesis alternation was proposed in (Hajnicz, 2011a). However, the information that two schemata are semantically related is useless unless we know which slots carry the same semantic information (represent the same argument). Each group of schemata forms a single semantic dictionary entry, represented as a list of semantic arguments. Slots of every schema are linked to corresponding semantic arguments.

2. Related works

There exist several manually prepared semantic valence dictionaries. For English, the most famous are FrameNet (Baker et al., 2003; Fillmore et al., 2003; Ruppenhofer et al., 2010), VerbNet (Dang et al., 1998; Kipper et al., 2000; Kipper-Schuler, 2005; Kipper et al., 2006) and PropBank (Kingsbury and Palmer, 2002; Kingsbury et al., 2002; Palmer et al., 2005). In the case of Slavic languages, two extensive dictionaries for Czech, Vallex (Žabokrtský and Lopatková, 2007) and VerbaLex (Hlaváčková and Horák, 2005, 2006) could be mentioned. Among these dictionaries, PropBank has abstract arguments ordered w.r.t. their obliqueness hierarchy (Keenan and Comrie, 1977, 1979; Croft, 2003), whereas VerbNet and VerbaLex have arguments interpreted by means of selectional preferences related to wordnet synsets.

3. Valence dictionary

A syntactic valence dictionary is a set of entries representing schemata for every verb considered. Each syntactic schema is a list of syntactic slots. The dictionary of 32 verbs chosen for the experiment was prepared on the basis of Świdziński's dictionary (Świdziński, 1994). Verbs were chosen manually in a way to maximise the variability of their syntactic schemata (in particular, diathesis alternations) on one hand and polysemy within a single schema on the other. Their frequency was an important criterion for this choice as well.

The list of slots can include: adjectival phrases (AdjP), adverbial phrases (AdvP), infinitival phrases (InfP), nominal phrases (NP), prepositional-adjectival phrases (PrepAdjP), prepositional-nominal phrases (PrepNP) and clauses (SentP). A special slot sie hosts the reflexive marker. Some slots are parametrised. The only parameter of AdjP and NP is their case, the only parameter of InfP is its aspect. PrepAdjP and PrepNP have two parameters: the form of the preposition and the case of its AdjP or NP complement, respectively. SentP has one parameter, namely the complementizer introducing the clause. Below we list syntactic dictionary entries for the verb *rozpocząć* (*to begin*).

(1) rozpocząć advp np:nom
rozpocząć np:acc np:inst np:nom
rozpocząć np:acc np:nom
rozpocząć np:acc np:nom prepnp:od:gen
rozpocząć np:acc np:nom prepnp:z:inst
rozpocząć np:inst np:nom sie
rozpocząć np:nom prepnp:dla:gen sie
rozpocząć np:nom prepnp:od:gen sie
rozpocząć np:nom sie

A syntactic-semantic valence dictionary was obtained by supplementing the syntactic valence dictionary with selectional preferences. Here, we consider the simple case of a fixed set of 25 semantic categories, which were assigned to nouns at the beginning of the preparation of

[1] We use the term syntactic *schema* instead of very popular syntactic *frame* in order to distinguish it from the term *semantic frame*.

the Polish WordNet (Piasecki et al., 2009), which was modelled on the Princeton WordNet (Fellbaum, 1998) and wordnets constructed in the EuroWordNet project (Vossen, 1998).

4. Classification of alternations

In (Hajnicz, 2011b) we presented a very coarse, purely syntactic classification of potential alternations, describing only how the alternations relate slots in two schemata involved. The alternations can be divided into two types. First, there are alternations which preserve the number of slots in both schemata. This condition is satisfied by alternations referred to in (Hajnicz, 2011b) as *simple* alternation, exemplified by dative alternation, see (2), *cross* alternation exemplified by locative alternation, see (3), *simple reflexive* alternation, cf. (4), and *cross-reflexive* alternation, cf. (5). The sub-examples differ in the animacy of arguments.

(2) *Chłopak posłał książkę koledze.*
(*A boy sent his friend a book.*)
Chłopak posłał książkę do kolegi.
(*A boy sent a book to his friend.*)

(3) a. *Trawa porosła wzgórze.*
(*Grass has grown over the hill.*)
Wzgórze porosło trawą.

b. *Kwiaty pachną w ogrodzie.*
(*Flowers smell in the garden.*)
Ogród pachnie kwiatami/od kwiatów.
(*The garden smells of flowers.*)

c. *Rolnik ładuje wóz sianem.*
(*The farmer loads the wagon with hay.*)
Rolnik ładuje siano na wóz.
(*The farmer loads hay onto the wagon.*)

(4) a. *Chłopiec ogania muchy /się od much.*
(*A boy drives away flies.*)

b. *Chłopak kocha dziewczynę/się w dziewczynie.*
(*A boy loves a girl.*)

(5) a. *Hrabina urodziła syna.*
(*A countess gave birth to a son.*)
Syn urodził się hrabinie.
(*A son was born to a countess.*)

b. *Córka niepokoi matkę.*
(*A daughter worries* (*her*) *mother.*)
Matka niepokoi się córką/o córkę.
(*Mother worries about* (*her*) *daughter.*)

c. *Temat interesuje badacza.*
(*The subject is interesting to the researcher.*)
Badacz interesuje się tematem.
(*The researcher is interested in the subject.*)

d. *Publiczność wypełniła teatr.*
(*The audience filled the theatre.*)
Teatr wypełnił się publicznością.
(*The theatre filled with the audience.*)

e. *Wino napełnia kieliszki.* (*Wine fills glasses.*)
Kieliszki napełniają się winem.
(*Glasses fill with wine.*)

f. *Kurz pokrył meble.*
(*Dust covered the furniture.*)
Meble pokryły się kurzem.
(*The furniture covered with dust.*)

Second, there are alternations in which one of the alternating slots is absent in one schema. This condition is satisfied by alternations referred to in (Hajnicz, 2011b) as *deletion* alternation exemplified by object drop alternation, see (6), and *shift* alternation exemplified by an unreflexive case of causative alternation, see (7). *Reflexive deletion* alternation is exemplified by reflexive alternation, cf. (8), and reciprocal alternation, cf. (9), where the reflexive marker *się* plays the role of *oneself* or *each other*, respectively.[2] *Reflexive shift* alternation is exemplified by causative alternation, cf. (10).

(6) *Matka pozmywała naczynia.*
(*Mother washed the dishes.*)
Matka pozmywała. (*Mother washed.*)

(7) a. *Jeździec pognał konia przez las.*
(*The rider rode a horse across a forest.*)
Koń pognał przez las.
(*A horse rode across a forest.*)

b. *Kelner napełnia kieliszki winem.*
(*The waiter fills glasses with wine.*)
Wino napełnia kieliszki. (*Wine fills glasses.*)

(8) *Żołnierz obronił towarzysza/się przed atakiem.*
(*A soldier defended his comrade/himself from the attack.*)

(9) *Chłopak spotkał dziewczynę / się z dziewczyną.*
(*A boy met a girl.*)
Chłopak i dziewczyna spotkali się (*ze sobą*).
(*A boy and a girl met* (*each other*).)

(10) a. *Kelner stłukł szklanki.*
(*A waiter broke glasses.*)
Szklanki stłukły się. (*Glasses broke.*)

b. *Kelner napełnia kieliszki winem.*
(*The waiter fills glasses with wine.*)
Kieliszki napełniają się winem.
(*Glasses fill with wine.*)

c. *Nadzorca zaharował niewolników* (*na śmierć*).
(*The overseer made slaves slog away* (*to death*).)
Niewolnicy zaharowali się (*na śmierć*).
(*Slaves slogged away* (*to death*).)

Semantically, *reflexive deletion* alternation preserves the number of slots, as *się* plays the role of the reflexive pronoun in the schema containing it carrying semantic information. Therefore, it is linked to a corresponding semantic argument in the semantic frame.

5. Rules of ordering slots

Our goal is to order verb slots in a group of schemata in a consistent way. Schemata in a group are semantically connected, which is accomplished by alternations relating them. The main idea is to find a dominant schema in

[2] These alternations cannot be differentiated at the level of schemata.

a group, order its slots and subsequently propagate this ordering to other schemata according to corresponding alternations.

Syntactic slots have the following priorities according to their obliqueness hierarchy, based on constructions as *Piotr przywiózł Annie kwiaty samochodem.* (*Peter brought Anna flowers by car.*, see also Ostler, 1979; Primus, 1999).

1. nominative np:nom,
2. accusative np:acc,
3. genitive np:gen,
4. dative np:dat,
5. instrumental np:inst,
6. prepositional phrases.

We have not employed any heuristics for ordering prepositional phrases not involved in alternations.

Each alternation contains information about which slots should share the same semantic argument; slots not involved in the alternation should agree both at the syntactic and semantic level. Thus, in order to find a dominant schema in the entire group we should establish the dominant schema in every pair participating in the particular alternation. For alternations "losing" slots, certainly a schema which contains all slots is dominant. Therefore, looking for the dominant schema, we only consider the longest ones as candidates.

Depending on the thorough analysis of language material, we formulated the following heuristics of domination of alternating schemata:

1. For simple and simple reflexive alternations, the schema in which the alternating slot has higher priority dominates;
2. For cross alternation,
   (a) if one of alternating slots is PrepNP, then the schema containing it dominates (cf. (3) b., c.),
   (b) otherwise a schema in which the alternating slot has higher priority dominates (cf. (3) a.);
3. For reflexive cross alternation, domination depends on the animacy of the active or reflexive subject
   (a) if the reflexive object is dative, then the active schema dominates, (cf. (5) a.),
   (b) if the reflexive subject is animate, then the reflexive schema dominates (cf. (5) b., c.),
   (c) if the reflexive subject is inanimate, then the active schema dominates (cf. (5) d.–f.),
   (d) if the subject is not involved in the alternation,[3] then a schema in which the alternating slot has higher priority dominates.

This procedure enables us to find the dominating schema in a group, order its slots and propagate this information to other schemata.

Some of considered slots are adjuncts "typically related to some verbs and not to others" (Žabokrtský and Lopatková, 2007). In PropBank such slots, usually prepositional, are put outside the obliqueness hierarchy and labelled as ArgM (modifiers). We follow this convention for prepositional phrases not involved in any alternation. Thus, slots having time, place, act or event as their strongest selectional preference are not ordered and labelled with M instead.

6. Experiments

The process of ordering slots is concurrent with the process of grouping schemata. The experiments were performed using the manually prepared set of alternations on the one hand (denoted as M) and the automatically obtained set of alternations (Hajnicz, 2011b, denoted as A). Alternations, in which the verb *rozpocząć* (*to begin*) participates, are presented in (12). Semantically consistent slots participating in the alternation are displayed as **np:nom**, whereas semantically dropped ones are displayed as np:nom. All schemata of the verb *rozpocząć* are semantically related and form a single group (11).

```
(11)  rozpocząć     1              A1, A2, A3
      np:acc:A2     np:inst:A3     np:nom:A1
      np:acc:A2     np:nom:A1
      np:acc:A2     np:nom:A1      prepnp:od:gen:A3
      np:inst:A3    np:nom:A2      sie
      np:nom:A2     prepnp:dla:gen:A1   sie
      np:nom:A2     prepnp:od:gen:A3    sie
      np:nom:A2     sie
```

Experiments were conducted in three ways. First, all slots were ordered. Secondly, two factors F of preference strength were considered: 1 and 1.5, meaning that "adjunctive" selection preference is F times greater than all other preferences in sum. As a baseline (denoted as B), ordering of slots was performed independently for each schema.

Evaluation was performed using a semantic dictionary prepared manually especially for our experiments. Its exemplary entry is presented in (11). The simplest way of evaluation would be to check for the exact match of labels. However, the results of grouping schemata could influence the label assignment. In order to evaluate label assignment exclusively, we decided to accept labelling as correct if ordering is preserved, i.e., all slots preceding and succeeding the validated one in the automatically obtained dictionary precede and succeed it in the manually prepared dictionary.

In spite of the above mentioned differences, each slot could be correctly labelled or not, there are no other possibilities. Thus, the only appropriate measure is correctness. We calculate it for particular slots and for the whole schemata, meaning that all their slots are properly ordered. Correctness was calculated in two ways. First, in a permissive way (denoted as ∪), in which modifier M agrees with any actual argument Ai. Secondly, in a rigorous way (denoted as ∩), in which all modifiers M must match.

The results of evaluation are presented in Table 1. Symbol ∞ represents the infinitive factor of the "adjunctive" preference strength, i.e., ordering all slots. In this case, only the permissive calculation of correctness is ap-

[3] Such cases of reflexive cross alternation were not found in the investigated set of verbs.

(12)

reflexive shift

np:acc np:inst np:nom	np:inst **np:nom** sie
np:acc np:nom	**np:nom** sie
np:acc np:nom prepnp:od:gen	**np:nom** prepnp:od:gen sie

simple

np:acc **np:inst** np:nom	np:acc np:nom **prepnp:od:gen**
np:inst np:nom sie	np:nom **prepnp:od:gen** sie

deletion

np:acc np:inst np:nom	np:acc np:nom
np:acc np:nom prepnp:od:gen	np:acc np:nom
np:inst np:nom sie	np:nom sie
np:nom prepnp:dla:gen sie	np:nom sie
np:nom prepnp:od:gen sie	np:nom sie

reflexive cross

np:nom prepnp:dla:gen sie	**np:acc np:nom**

Table 1: Evaluation of labelling slots in a semantic dictionary

data	slots			schemata		
	∞	1	1.5	∞	1	1.5
$M^{\cup}$	99.89	99.89	99.89	99.76	99.76	99.76
$M^{\cap}$	99.42	99.29	99.77	98.81	98.57	99.52
$A^{\cup}$	97.44	97.69	97.69	97.21	97.44	97.44
$A^{\cap}$	96.63	97.09	97.24	95.81	96.51	96.74
$B^{\cup}$	96.93	98.69	97.30	96.39	98.19	96.75
$B^{\cap}$	96.12	80.08	94.27	94.95	70.40	90.97

propriate, while the rigorous one is presented for the sake of comparison.

The results of evaluation are impressive, being good even for the baseline. Observe that factor 1 considers too many slots as modifiers, which results in the worst restrictive evaluation. Thus, factor 1.5 should be considered optimal. Permissive evaluation depends on the choice of modifiers rather weakly, only for the automatically detected alternations. This means that modifiers only mask errors in the numbering of slots, which manifests itself in a difference between permissive and restrictive evaluation, greater than for manually prepared alternations.

Results for automatically detected alternations are 2-3 percentage points worse than results for manually prepared alternations. This means that the method is sensitive to data errors. Still, they are better than the baseline, especially if modifiers are detected. The reason is that the baseline method considers all prepositional phrases as potential modifiers, having no information about their participation.

7. Conclusions and future work

In this paper, a method of consistent semantic labelling of slots in a group of semantically related verb schemata was proposed. It is based on several heuristics concerning the obliqueness hierarchy of slots in a schema and domination of schemata related by a particular diathesis alternation.

The experiments performed on a set of verbs larger than the present one of 32 verbs would give more reliable results. However, the chosen verbs have proportionally large sets of related schemata, hence the task of ordering their slots is probably harder than in average.

The results are so good that the method could be applied fully automatically, especially for manually prepared or corrected sets of alternations. This by no means concerns other steps of creating a semantic valence dictionary. Thus, future efforts should be devoted to the improvement of methods of alternation detection and grouping schemata. Nevertheless, the whole process could be performed only semi-automatically. The automatic detection of selectional preferences, alternations and automatic grouping of schemata will be a valuable support for lexicographers creating such a dictionary.

(13) rozpocząć 1

A1	person–0.8,	group–0.2	
A2	act–0.45,	event–0.3,	time–0.1
A3	event–0.4,	act–0.3,	artifact–0.2

Semantic dictionary entries, like (11), representing particular verb meanings, are assigned to sets of semantically related schemata of verbs, with the uniform numbering of slots. Such an entry, together with selectional preferences of arguments, as in (13), could be a valuable source of information about their semantic roles. This is an interesting subject for further research as well.

Acknowledgements This research is supported by the POIG.01.01.02-14-013/09 project which is co-financed by the European Union under the European Regional Development Fund.

References

Baker, Colin F., Charles J. Fillmore, and Beau Cronin, 2003. The structure of the FrameNet database. *International Journal of Lexicography*, 16(3):281–296.

Croft, William, 2003. *Typology and Universals*. Cambridge, Great Britain: Cambridge University Press.

Dang, Hoa Trang, Karin Kipper, Martha Palmer, and Joseph Rosenzweig, 1998. Investigating regular sense extensions based on intersective Levin classes. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational LinguisticsCOLING-ACL'98*. Montreal, Canada.

Fellbaum, Christiane (ed.), 1998. *WordNet — An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Fillmore, Charles J., Christopher R. Johnson, and Miriam R. L. Petruck, 2003. Background to FrameNet. *International Journal of Lexicography*, 16(3):235–250.

Hajnicz, Elżbieta, 2009a. Generalizing the EM-based semantic category annotation of NP/PP heads to wordnet synsets. In Zygmunt Vetulani (ed.), *Proceedings of the 4th Language & Technology Conference*. Poznań, Poland.

Hajnicz, Elżbieta, 2009b. Problems with pruning in automatic creation of semantic valence dictionary for Polish. In Václav Matoušek and Pavel Mautner (eds.), *Proceedings of the International Conference on Text, Speech and Dialogue TSD 2009*, volume 5729 of *LNAI*. Pilsen, Czech Republic: Springer-Verlag.

Hajnicz, Elżbieta, 2009c. Semantic annotation of verb arguments in shallow parsed Polish sentences by means of EM selection algorithm. In Małgorzata Marciniak and Agnieszka Mykowiecka (eds.), *Aspects of Natural Language Processing*, volume 5070 of *LNCS*. Springer-Verlag, pages 211–240.

Hajnicz, Elżbieta, 2010. Aggregating entries of semantic valence dictionary of Polish verbs. In Pier Marco Bertinetto, Anna Korhonen, Alessandro Lenci, Alissa Melinger, Sabine Schulte im Walde, and Aline Villavicencio (eds.), *Proceedings of the Interdisciplinary Workshop on the Identification and Representation of Verb Features (Verb 2010)*. Pisa, Italy: Scuola Normale Superiore and Università di Pisa.

Hajnicz, Elżbieta, 2011a. Grouping alternating schemata in semantic valence dictionary of Polish verbs. In Václav Matoušek, XXX, and XXX (eds.), *Proceedings of the International Conference on Text, Speech and Dialogue TSD 2011*. Pilsen, Czech Republic: Springer-Verlag.

Hajnicz, Elżbieta, 2011b. Similarity-based method of detecting diathesis alternations in semantic valence dictionary of Polish verbs. In P. Bouvry, Mieczysław A. Kłopotek, F. Leprevost, Małgorzata Marciniak, Agnieszka Mykowiecka, and H. Rybiński (eds.), *International Joint Conference on Security and Intelligent Information Systems*, LNAI. Warsaw, Poland.

Hlaváčková, Dana and Aleš Horák, 2005. Transformation of WordNet Czech valency frames into augmented VALLEX-1.0 format. In Zygmunt Vetulani (ed.), *Proceedings of the 2nd Language & Technology Conference*. Poznań, Poland.

Hlaváčková, Dana and Aleš Horák, 2006. VerbaLex — new comprehensive lexicon of verb valences for Czech. In *Proceedings of the Third International Seminar on Computer Treatment of Slavic and East European Languages*. Bratislava, Slovakia.

Keenan, Edward L. and Bernard Comrie, 1977. Noun phrase accesibility and universal grammar. *Linguistic Inquiry*, 8(1):63–98.

Keenan, Edward L. and Bernard Comrie, 1979. Noun phrase accesibility and universal grammar revisited. *Language*, 55(3):649–664.

Kingsbury, Paul and Martha Palmer, 2002. From TreeBank to PropBank. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002)*. Las Palmas, Spain.

Kingsbury, Paul, Martha Palmer, and Mitchell P. Marcus, 2002. Adding semantic annotation to the Penn TreeBank. In *Proceedings of the Human Language Technology Conference*. San Diego, CA.

Kipper, Karin, Hoa Trang Dang, and Martha Palmer, 2000. Class-based construction of a verb lexicon. In *Proceedings of the 17th National Conference on Artificial Intelligence*. Austin, TX: AAAI Press.

Kipper, Karin, Anna Korhonen, Neville Ryant, and Martha Palmer, 2006. Extending VerbNet with novel verb classes. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-2006)*. Genoa, Italy.

Kipper-Schuler, Karin, 2005. *VerbNet: A broad coverage, comprehensive verb lexicon*. PhD thesis, Computer and Information Science Department, University of Pennsylvania.

Ostler, Nicholas David MacLachlan, 1979. *Case-linking: A Theory of Case and Verb Diathesis Applied to Classical Sanskrit*. PhD thesis, Massachussets Institute of Technology, Cambridge, MA.

Palmer, Martha, Paul Kingsbury, and Daniel J. Gildea, 2005. The proposition bank: an annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Piasecki, Maciej, Stanisław Szpakowicz, and Bartosz Broda, 2009. *A Wordnet from the Ground Up*. Wrocław, Poland: Oficyna Wydawnicza Politechniki Wrocławskiej.

Primus, Beatrice, 1999. *Cases and Thematic Roles: Ergative, Accusative and Active*. Tübingen, Germany: Max Niemeyer Verlag.

Ruppenhofer, Josef, Michael Ellsworth, Miriam R. L. Petruck, and Christopher R. Johnson, 2010. FrameNet II: Extended theory and practice. Internet.

Świdziński, Marek, 1994. *Syntactic Dictionary of Polish Verbs*. Uniwersytet Warszawski / Universiteit van Amsterdam.

Vossen, Piek (ed.), 1998. *EuroWordNet: a multilingual database with lexical semantic network*. Dordrecht, Holland: Kluwer Academic Publishers.

Žabokrtský, Zdeněk and Markéta Lopatková, 2007. Valency information in VALLEX 2.0: Logical structure of the lexicon. *The Prague Bulletin of Mathematical Linguistics*, 87:41–60.